DECREASED RESPONSE TIME FOR PPRC WRITE OPERATION

TECHNICAL FIELD

[1] The present invention relates generally to peer-to-peer remote copy (PPRC) storage systems and, in particular, to reducing the number of round trips required to complete a write operation between a primary storage control unit and a secondary storage control unit.

BACKGROUND ART

[2] Data integrity and availability is a critical factor in large computer data systems. Consequently, backup data storage systems have been developed to prevent the loss of data in the event of various types of failures. One such backup system is known as "peer-to-peer remote copy" (PPRC). As illustrated in Fig. 1, in a PPRC system 100, data generated by a host device 110 is stored on a primary storage unit 120. A copy of the data is also transmitted, such as over a fibre channel network 130, and stored on a secondary storage unit 140. Because of the flexibility of network interconnections, the primary and secondary units 120 and 140 may be physically located remote from the host 110. And, for data security, the primary and secondary units 120 and 140 are physically located distant from each other, thereby reducing the likelihood of a single disaster simultaneously harming both the primary and secondary units 120 and 140.

[3] The distance by which the primary and secondary units 120 and 140 may be separated is dependent upon numerous factors. One significant factor is the total response time of each I/O operation (such as a write operation); that is, the amount of time required for a block of data to be transferred from the primary storage unit 120 to the secondary storage unit 140, including all handshaking. Typically, the longer the response time, the shorter the distance which may practically separate the two units. And, a significant factor in determining the response time is the number of round trips of command and data which must take place to complete a transfer of data. As will be appreciated, the more round trips which are necessary, the slower the effective transfer rate becomes.

Docket: TUC920030144US1
Express Mail Label: EV303489215US

[4] One such round trip occurs when the primary and secondary units 120 and 140 exchange "transfer ready" signals prior to a write operation. The primary unit 120 (also known as the initiator) transmits a message to the secondary unit 140 (also known as the target) indicating that data is ready to be transferred. Until the primary unit 120 receives an appropriate acknowledgement from the secondary unit 140, transfer of the data cannot begin. Among other items, the acknowledgement indicates that the secondary unit has prepared the necessary buffers and is ready to receive the data. Such preparation may entail some delay and the handshaking itself results in some delay as well. Thus, such a transfer requires two round trips (the transfer ready exchange and the transfer of data with a subsequent acknowledgement of receipt) and results in a corresponding delay.

[5] Another round trip occurs when additional control information is transferred from the primary unit 120 to the secondary unit 140 before the data itself is transferred. Such additional control information may not be able to fit within a conventional write command, such as a command descriptor block (CDB). Thus, another round trip is necessary to separately transfer the control information.

[6] Other factors may necessitate further round trips. Transferring data over a conventional fibre channel network may entail three to four round trips and the distance between the primary storage unit 120 and the secondary storage unit 140 may thus be limited to about 100 kilometers.

[7] Consequently, in order to increase the distance between the primary and secondary units, it remains desirable to reduce the response time for data transfers.

SUMMARY OF THE INVENTION

[8] The present invention provides method, system and computer program product to improve the efficiency of data transfers in a PPRC environment. Any or all of three features may be implemented, each of which reduces the number of round trips required for the exchange of handshaking, data and control information. A first feature includes disabling the "transfer ready" acknowledgment which normally occurs between a primary storage controller and a secondary storage controller. Disabling such acknowledgement eliminates one round trip.

[9] A second feature includes pre-allocating payload and data buffers in the secondary storage controller. Pre-allocating buffers permits buffers to be available immediately upon receipt of payload information and data blocks, thereby reducing response time. A third feature includes packaging write control information with a write command in an extended command descriptor block (CDB). Such a step eliminated the need for a separate transmission of the write control information. The CDB is transmitted along with a data block from the primary storage controller to the secondary storage controller and placed in the respective, pre-allocated buffers.

[10] Data may be pipelined to the secondary; that is, a block of data may be transmitted from the primary without waiting for acknowledgment from the secondary that the previous block of data was received.

[11] By increasing the response time for data transfers, the distance separating the primary and secondary storage controllers may be increased.

BRIEF DESCRIPTION OF THE DRAWINGS

[12] Fig. 1 is a block diagram of an exemplary PPRC data storage system;

[13] Fig. 2 is a block diagram of a data storage system in which the present invention may be implemented;

[14] Fig. 3 is a flow chart of a method of the present invention;

[15] Fig. 4 illustrates an exemplary command descriptor block;

[16] Fig. 5A illustrates an exemplary extended command descriptor block; and

[17] Fig. 5B illustrates an exemplary data block.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[18] Fig. 2 is a block diagram of a data storage system 200 in which the present invention may be implemented. The system 200 may be a PPRC system in which a host device 210 transmits data to a primary storage controller 220 for storage on a storage device (such as an array of disk dives) 228. Additionally, the primary storage controller 220 transfers a copy of the data over a network 230 to a remote secondary storage controller 240 for storage on a storage device 248. The primary storage controller 220 further includes a processor 222 for executing instructions

relating to the transfer of the copies of data to the secondary storage controller 240. The secondary storage controller also includes memory which may be allocated to payload buffers 244 to hold incoming command and control information and data buffers 246 to hold incoming data. A processor 242 in the secondary storage controller 240 executes instructions relating to the allocation of memory space to buffers and to the receipt and ultimate storage of data.

[19] Referring to the flow chart of Fig. 3, in operation, during a log-on process (step 300), the processors 222 and 242 in the primary and secondary storage controllers 220 and 240 optionally disable the "transfer ready" acknowledge requirement (step 302) and exchange acknowledgements to that effect (step 304). This step, if taken, eliminates one round trip of handshaking between the two units. The secondary controller 240 is then directed to pre-allocate some of its memory to payload buffers 244 (step 306). Each payload buffer is of sufficient size to hold a PPRC fibre channel protocol (FCP) command containing write control information as well as the write command itself. As will be described below, both data structures may be "packaged" in an extended command descriptor block (CDB) to eliminate another round trip, one which is not dedicated to the transfer of actual data to the secondary storage controller 240. The secondary controller 240 is also directed to pre-allocate some of the memory to data buffers 246 (step 308), each sufficient in size to hold a block of PPRC write data. The number of buffers of each type which are pre-allocated (that is, set aside prior to a data block actually being received by the controller) is not critical but may be determined by balancing performance (leading to more buffers being pre-allocated) against cost and available memory (leading to fewer buffers being pre-allocated).

[20] The primary and secondary controllers 220 and 240 also establish the size of data which may be transferred in one write operation (step 310). Herein, the unit of data to be transferred will be referred to as a "block".

[21] When the primary storage controller 220 issues a PPRC write command to commence a write operation (step 312), the primary storage controller 220 uses an extended CDB 500 (Fig. 5A) (step 314) to contain write control instructions 502 as well as the write command 504 to be transmitted to the secondary storage controller

240. A conventional CDB 400 (Fig. 4) is used to transmit only the write command 404 and has a size of approximately 16 bytes. In order to decrease the response time of a write operation and increase the efficiency of data transfers, an extended CDB 500, of approximately 80 to 96 bytes, is employed. The extended CDB 500 is carried inside the FCP command which is transmitted by the primary storage controller 220 to the secondary storage controller 240 (step 316). Subsequently, and without having received any "ready" signal from the secondary storage controller 240, the primary storage controller 220 transmits a block of data 510 (Fig. 5B) to the secondary storage controller 240 (step 318).

[22] Upon receiving the FCP command from the primary storage controller 220 (step 320), the processor 242 of the secondary storage controller 240 directs that the FCP command be placed in one of the payload buffers 244 (step 322). The block of data 510 is received (step 324) and placed in one of the data buffers 246 (step 326). The contents of both the payload buffer 244 and data buffer 246 then are processed by the PPRC application executing inside the secondary storage controller 240 (step 328). The block of data 510 is ultimately stored on the storage device 248 (step 330), thereby completing the write operation. When the write operation has been completed, the secondary storage controller 240 transmits a status signal back to the primary storage controller 220 (step 332) and closes the exchange (step 334). The secondary storage controller 240 then releases the payload and data buffers (step 336) for subsequent re-use. When the primary storage controller 220 receives the status signal (step 338), it too closes the exchange (step 340).

[23] Because the present invention supports "piping" of data, prior to the primary controller 220 receiving the status signal, the primary controller 220 may prepare a second FCP command, with write control information and a write command packaged in a second extended CDB (step 342). The second FCP command and an associated block of data may then be transmitted to the secondary controller 340 (step 344). Upon receipt by the secondary controller 240 (step 346), the second FCP command is placed in a payload buffer (step 348) and the second block of data is placed in a data buffer (step 350). The buffer in which the second FCP command is placed may be a second payload buffer (different from the first payload buffer) or

may be the first payload buffer (if the first payload buffer has previously been released). Moreover, if the first payload buffer has not previously been released and no other payload buffer is available (that is, was not pre-allocated), the processor 242 may allocate additional memory space as an additional payload buffer, available to receive the second FCP command. Similarly, the second block of data is received into a data buffer, which may be the first data buffer (after being released), another pre-allocated data buffer or a newly allocated data buffer.

[24] Each of the three steps alone, disabling the transfer ready acknowledgement, pre-allocating buffers, and packaging write control information with a write command in an extended CDB, reduces the number of round trips required for the transfer of data. Together, the number of required round trips may be reduced to one, thereby significantly decreasing the response time of a write operation, increasing total throughput and/or increasing the possible distance by which the primary and secondary storage controller 220 and 240 may be separated. In fact, the separation distance may be increased to approximately 400 km.

[25] The objects of the invention have been fully realized through the embodiments disclosed herein. Those skilled in the art will appreciate that the various aspects of the invention may be achieved through different embodiments without departing from the essential function of the invention. The particular embodiments are illustrative and not meant to limit the scope of the invention as set forth in the following claims.